revista.tradumàtica
tecnologies de la traducció

150-156

Motivos del creciente uso de traducción automática seguida de posedición

Felipe Sánchez-Martínez
Grup Transducens
Dept. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
fsanchez@dlsi.ua.es

RESUMEN

Este artículo aborda las causas que, en opinión del autor, han motivado la adopción creciente de sistemas de traducción automática (TA) para la producción de borradores para la posedición. Estas causas son principalmente cuatro: la mejora en las técnicas de TA, la mayor disponibilidad de recursos tales como software y datos, el cambio en las expectativas de los usuarios en cuanto a lo que se puede esperar o no de un sistema de TA, y por último la mayor integración de sistemas TA en entornos de ayuda a la traducción.

Palabras clave: traducción automática, evolución, posedición, diseminación

RESUM *(Motius del creixent ús de la traducció automàtica seguida de postedició)*

Aquest article aborda les causes que, en opinió de l'autor, han motivat l'adopció crecient de sistemes de traducció automàtica (TA) per produir esborranys per a la postedició. Aquestes causes són principalment quatre: la millora en les tècniques de TA, la major disponibilitat de recursos com ara programari i dades, el canvi en les expectatives dels usuaris quant al que es pot esperar o no d'un sistema de TA, i finalment la major integració de sistemes TA en entorns d'ajuda a la traducció.

Paraules clau: traducció automàtica, evolució, postedició, disseminació

ABSTRACT *(Reasons for the increasing use of Machine translation followed by post-editing)*

This article discusses the causes which, in the author's opinion, have led to an increase of the adoption of machine translation (MT) to produce drafts for post-editing. There are four main causes for this: improvement of MT techniques, increased availability of resources such as software and data, a change in users' expectation about MT, i.e. what can and cannot be expected from an MT system, and better ways of integrating MT systems in compute-aided translation tools.

Keywords: machine translation, evolution, post-editing, dissemination

1. Introducción

El uso de sistemas de traducción automática (TA) para la producción de borradores para la posedición ha crecido en los últimos años. Así lo atestigua el informe publicado por TAUS (2009) en el que se estudiaron las prácticas en lo que respecta a la automatización del proceso de traducción de varios proveedores de servicios lingüísticos: de estos, el 40% declaró hacer uso de TA, mientras que del 60% restante, el 89% dijo tener planes para la incorporación de TA en los próximos dos años. El motivo de este incremento en la adopción de sistemas de TA es la ganancia de productividad que se obtiene, como se desprende de diversos estudios (Guerberof, 2009; de Almeida & O'Brien, 2010; Plitt &Masselot, 2010). De acuerdo con TAUS (2009) esta ganancia está en torno al 70%, con una reducción de costes de entre el 30% y el 40%.

Este artículo aborda las causas que, en opinión del autor, han motivado que hoy en día la TA sea una tecnología útil que permite ahorrar costes. El artículo se centra en cuatro causas principales: la mejora de las técnicas de TA, la mayor disponibilidad de recursos y herramientas para el desarrollo de sistemas de TA, el cambio en las expectativas de los usuarios de la TA, y por último, la mayor integración de sistemas de TA en herramientas de ayuda a la traducción. Cabe decir que estas no son las únicas causas que han motivado el mayor uso de TA para la diseminación; algunas de las causas no discutidas en este artículo son la disponibilidad de sistemas de TA en línea y sin coste alguno, el aumento en la velocidad de traducción o el mayor alfabetismo digital de los traductores.

El resto del artículo se organiza como sigue. El siguiente apartado revisa brevemente las distintas aproximaciones a la TA y ofrece una descripción de las mejoras que estas técnicas han sufrido en los últimos años. El apartado 3, describe los recursos disponibles en Internet, tanto programas como datos y herramientas, que han facilitado el desarrollo de sistemas de TA por parte de empresas e instituciones y que por tanto también ha tenido una gran incidencia en la adopción de sistemas de TA para la producción de borradores para la pos-edición. Los apartados 4 y 5, por su parte, tratan el cambio en las expectativas de los usuarios y la integración de sistemas de TA en entornos de ayuda a la traducción, respectivamente. El artículo termina con unas conclusiones a modo de resumen.

2. Técnicas de traducción automática

Los sistemas de TA pueden dividirse en función de la aproximación seguida para llevar a cabo la traducción entre sistemas basados en conocimiento y sistemas basados en corpus. Los sistemas basados en conocimiento, por lo general sistemas de TA indirecta por transferencia sintáctica (Hutchins & Somers, 1992), aplican transformaciones (reglas) a la representación intermedia del texto en lengua origen que se obtiene tras el análisis del mismo para convertirla en una representación intermedia de la lengua destino a partir de la cual se genera la traducción. Tanto para el análisis del texto en lengua origen como para la generación del texto en lengua meta es necesario contar con herramientas específicas tales como analizadores morfológicos o sintácticos. Si bien es cierto que desde la concepción de esta aproximación a la TA ha habido avances en algunas de las técnicas empleadas por parte de estas herramientas, no puede decirse lo mismo en cuanto a las técnicas básicas de TA basadas en conocimiento: la arquitectura de los sistemas actuales es muy similar a la de los primeros sistemas de TA.

En cuanto a los sistemas basados en corpus cabe diferenciar dos aproximaciones distintas: TA basada en ejemplos (Carl & Way, 2003) y TA estadística (Koehn, 2010), ambas basadas en la disponibilidad de textos paralelos (frases en lengua origen junto con su traducción a la lengua destino) en cantidad suficiente. En TA basada en ejemplos, la traducción se realiza por analogía: dados uno o más textos paralelos, el sistema analiza la frase a traducir, la divide en segmentos más pequeños cuyas traducciones son recuperadas de los textos paralelos y las combina para producir una nueva traducción en lengua meta. En

la actualidad los sistemas de TA basada en ejemplos, tales como Cunei (Phillips, 2011) o Matrex (Penkale et al., 2010), son sistemas híbridos con una fuerte componente de TA estadística por lo que en lo sucesivo serán considerados junto con este último tipo de sistemas de TA.

Los sistemas de TA estadística, el paradigma predominante en TA basada en corpus, utilizan para llevar a cabo la traducción modelos estadísticos cuyos parámetros se aprenden de forma automática a partir de textos paralelos y monolingües; la principal diferencia con respecto a los sistemas de TA basados en ejemplos radica en la forma de llevar a cabo la traducción, combinando varios modelos estadísticos para puntuar las traducciones, y en que una vez aprendidos los modelos estadísticos usados para traducir, los textos usados durante el aprendizaje son desechados. Entre los modelos estadísticos que se utilizan cabe destacar el modelo de traducción que relaciona unidades de ambas lenguas y el modelo de lengua que sirve para evaluar la fluidez o naturalidad de las traducciones.1

Los sistemas de TA estadística han experimentado diversos cambios desde su concepción a finales de la década de los 80 y principios de los 90. Los primeros sistemas de TA estadística usaban la palabra como unidad mínima de traducción (Brown et al., 1993), de modo que el modelo de traducción no tenía en cuenta el contexto local a la hora de traducir, lo que provocaba, entre otros problemas, selecciones léxicas erróneas. Estos problemas se solventaron en parte con la introducción de los sistemas de TA estadística basada en segmentos bilingües (bilingual phrases en la literatura en inglés; Koehn, Och & Marcu, 2003), los cuales usan como unidad mínima de traducción segmentos de longitud variable. Estos segmentos no tienen por qué corresponder a unidades lingüísticas y permiten contemplar el contexto local dentro del propio segmento. Por lo general, los sistemas de TA estadística basados en segmentos bilingües tienen problemas a la hora de realizar reordenamientos o concordancias de larga distancia. Para paliar este problema se introdujeron los sistemas de TA estadística jerárquicos basados en segmentos bilingües (Chiang, 2007). Estos sistemas también utilizan segmentos bilingües de longitud variable como unidad mínima de traducción, pero con la novedad de permitir huecos dentro de estos segmentos, la traducción de los cuales se obtiene de otros segmentos bilingües que a su vez pueden contener huecos. La aparición de sistemas jerárquicos de TA estadística ha facilitado la introducción de sintaxis en la TA estadística de forma que la traducción puede hacerse entre árboles de análisis sintáctico de forma similar a como lo hacen los sistema de TA basados en conocimiento, pero en los que la obtención de las reglas de transferencia se hace de forma automática a partir de textos paralelos y su aplicación viene determinada por modelos estadísticos.

3. Disponibilidad de recursos

Independientemente de la aproximación a la TA elegida, un aspecto clave para el éxito de un sistema de TA son los recursos que este utiliza. Estos recursos pueden dividirse en recursos hardware, software y datos. En lo que respecta a los recursos hardware es obvio el abaratamiento de los mismos y el aumento de su potencia de cálculo, un aspecto clave cuando se trata de aprender los modelos estadísticos usados en TA estadística a partir de una gran colección de textos paralelos.

En cuanto a los recursos software, es decir, la implementación de aplicaciones que lleven a cabo la traducción de forma automática, se han liberado en los últimos años un número creciente de programas libres/de código fuente abierto,2 tales como la plataforma de TA basada en reglas Apertium (Forcada et al., 2011) o el sistema de TA estadística Moses

1 Véase la guía para lingüistas y traductores sobre TA estadística de Hearne & Way (2011) para una explicación detalla de cómo funcionan los sistemas de TA estadística basada en segmentos bilingües, el tipo de sistemas de TA estadística predominante en la actualidad.

2 Véase la definición de software libre/de código fuente abierto proporcionada por el proyecto GNU: http://www.gnu.org/philosophy/free-sw.es.html (última visita: 23 de enero de 2013)

(Koehn et al., 2007).3 La existencia de programas de TA libres/de código fuente abierto ha favorecido la adopción de sistemas de TA sin necesidad de invertir grandes cantidades de dinero en implementar técnicas y métodos de TA que han demostrado su utilidad en el ámbito experimental. La gran mayoría de estos programas han sido desarrollados por la comunidad científica y se encuentran en constante desarrollo, incorporando los últimos avances científicos. De este modo, instituciones y empresas se benefician de forma directa de los últimos avances en investigación, además de poder modificar los programas para ajustarlos a las necesidades concretas de su organización, una de las libertades básicas que garantizan las licencias de software libre/de código fuente abierto.4

Además de recursos hardware y software, para llevar a cabo la TA se requieren datos para el par de lenguas en cuestión. Estos datos incluyen, entre otros, diccionarios morfológicos y bilingües, reglas de desambiguación y reglas de transferencia estructural para el caso de los sistemas de TA basados en conocimiento, y textos paralelos y monolingües en cantidad suficiente para el aprendizaje de los modelos estadísticos en el caso de sistemas de TA estadística. La existencia de datos lingüísticos abiertos (diccionarios y reglas) codificados en formatos abiertos que favorecen su aprovechamiento por distintos sistemas ha facilitado el desarrollo de nuevos sistemas de TA basados en conocimiento para pares de lenguas con pocos recursos (Armentano-Oller & Forcada, 2008). Por su parte, la existencia de grandes colecciones de textos paralelos como Europarl (Koehn, 2005), el corpus de las Naciones Unidas (Rafalovitch & Dale, 2009) o los textos paralelos recopilados en el proyecto OPUS (Tiedemann, 2009), de fuentes tan diversas como la Agencia Europea del Medicamento o subtítulos de películas, han facilitado el desarrollo de sistemas de TA estadística. Estos corpus están en continuo crecimiento, lo que permite una mejora constante, aunque pequeña, de la calidad de las traducciones producidas por los sistemas que aprenden de ellos; en concreto, cada vez que se duplica la cantidad de texto paralelo usado para el aprendizaje de los modelos estadísticos usado para la traducción, se obtiene un incremento de aproximadamente un 2,5% en la medida de evaluación automática de la calidad de la traducción más utilizada en TA estadística (Och, 2005).5

Pese a la existencia de datos lingüísticos y textos paralelos disponibles en Internet, cabe decir que el aspecto que mayor incidencia tiene, junto con la disponibilidad de software libre/de código fuente abierto, en la calidad de los borradores producidos por los sistemas de TA actuales es el hecho de disponer, por parte de empresas e instituciones, de recursos propios tales como textos paralelos de la misma temática que los textos nuevos a traducir, pero también glosarios, para el entrenamiento de sistemas de TA estadística o para la adaptación de los diccionarios usados por los sistemas de TA basados en conocimiento.

4. Expectativas de los usuarios

Además de por el incremento en la calidad de las traducciones por los motivos expuestos más arriba, el uso de sistemas de TA para la producción de borradores para su posedición ha venido también motivado por un cambio en la expectativas de los usuarios con respecto a la calidad de las traducciones y su utilidad. En un principio, se esperaba que los traductores automáticos produjeran traducciones de calidad similar a la de las traducciones producidas por humanos, motivo por el cual la TA era descartada, dada su inhabilidad para producir

3 Véase el número especial de la revista Machine Translation (volumen 25, número 2) para una descripción más detalla de algunos de los sistemas de TA de código fuente libre/abierto disponibles en Internet (Sánchez-Martínez & Forcada, 2011). Una lista de sistemas de TA de código fuente libre/abierto se encuentra disponible en http://www.fosmt.org (última visita: 23 de enero de 2013)

4 Una lista de licencias de software libre/de código fuente abierto se puede consultar en http://opensource.org/licenses (última visita: 23 de enero de 2013)

5 Debe tenerse en cuenta que la mejora de la calidad de las traducciones medida de forma automática no necesariamente implica una mejora real que redunde en una reducción de la necesidad de posedición de las traducciones resultantes.

traducciones que parecieran humanas. Sin embargo, en la actualidad se ha demostrado que la TA, pese a ser una tecnología que en ocasiones produce traducciones ininteligibles, en muchos casos proporciona traducciones cuya posedición se puede realizar en un tiempo inferior al que requeriría realizar la traducción completamente desde cero. Dicho esto, cabe destacar que todavía hoy existe una percepción errónea por parte de algunos traductores en lo que respecta al impacto que el uso de TA tiene sobre su trabajo.

De entre los distintos estudios sobre la productividad de la posedición citados en la introducción, merece la pena destacar el estudio llevado a cabo por Plitt & Masselot (2010) sobre el incremento de productividad que se produce como resultado de usar TA para la localización de software. De los datos proporcionados por este estudio, ampliado recientemente y cuyos últimos resultados se encuentran disponibles en Internet,[6] se pueden extraer dos conclusiones interesantes. La primera es que, en todos los casos, traductores y pares de lenguas (en el estudio se traduce de inglés a nueve lenguas distintas), se produce un incremento de productividad como resultado de usar un sistema de TA estadística entrenado con textos paralelos de la propia organización. La segunda conclusión es que la percepción que los traductores tienen de su propia productividad con respecto al uso o no de TA no siempre es acertada; mientras que todos los traductores demostraron ser más productivos usando TA seguida de posedición, muchos de ellos tenían la percepción contraria, afirmando que eran más productivos cuando traducían usando un sistema de ayuda a la traducción basado en memorias de traducción.

5. Integración en sistemas de ayuda a la traducción

El último aspecto que, en opinión del autor, ha incidido en un mayor uso de TA seguida de posedición es la integración de sistemas de TA en la estación de trabajo habitual del traductor, las herramientas de ayuda a la traducción basadas en memorias de traducción. Aplicaciones como DéjàVu, SDL Trados, ESTeam o el programa libre/de código fuente abierto OmegaT incorporan TA (Lagoudaki, 2008; Cocci, 2009), ya sea simplemente para traducir aquellos segmentos para los cuales no se ha encontrado ninguna coincidencia parcial en la memoria de traducción, o para ayudar a traducir aquellas partes que no coinciden con el segmento en lengua origen de la unidad de traducción que se esté reutilizando. DéjàVu integra un sistema de TA basado en ejemplos para sugerir traducciones en aquellos casos en que no hay coincidencias exactas pero sí coincidencias parciales a partir de las cuales construir la propuesta de traducción (AutoAssemble; Garcia, 2005, Lagoudaki, 2008); ESTeam identifica las partes del segmento en lengua destino a modificar y usa TA para proponer traducciones para las mismas (Kranias & Samiotou, 2004); OmegaT simplemente ofrece la posibilidad de usar TA para traducir de formar completa el segmento con el que se esté trabajando. En cualquier caso, facilitar el uso de TA desde la estación de trabajo habitual del traductor le permite combinar las distintas tecnologías de la traducción a su alcance, en lugar de tener que elegir entre ellas, y obtener así el máximo provecho de las mismas.

6. Conclusión

A lo largo de este artículo se han abordado las causas que a juicio del autor han tenido una mayor incidencia en la adopción por parte de empresas, instituciones y particulares de sistemas de TA para la producción de borradores para su posedición y posterior publicación. Cuatro son las causas principales: mejoras en las técnicas de TA, mayor disponibilidad de recursos para el desarrollo de sistemas de TA, el cambio en las expectativas de los usuarios y por último una mayor integración de sistemas de TA en la estación de trabajo del traductor.

6 http://translate.autodesk.com/productivity.html (última visita: 23 de enero de 2013)

De todas ellas merece especial mención el aumento de recursos tales como el software libre/de código fuente abierto y la disponibilidad por parte de empresas e instituciones de colecciones de textos paralelos que permitan entrenar sistemas de TA estadística para la traducción de textos de la misma naturaleza que los usados durante el entrenamiento, o de glosarios para la adaptación de los diccionarios usados por los sistemas de TA basados en conocimiento. De cara al futuro se espera que la adopción de sistemas de TA continúe en aumento por los motivos anteriormente expuestos y como resultado de nuevas formas de integrar la TA en sistemas de ayuda a la traducción.

References

Armentano-Oller, C., M. L. Forcada (2008). "Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas", Procesamiento del Lenguaje Natural, 41:243-250.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer (1993). "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics, 19(2):263-311.

Carl, M., A. Way, editores (2003). "Recent advances in example-based machine translation", Text, Speech and Language Technology 21, Kluwer Academic Publishers.

Chiang, D. (2007). "Hierarchical phrase-based translation", Computational Linguistics, 33(2):201-228.

Cocci, L. (2009). "CAT tools for beginners", Translation Journal 13(4).

de Almeida, G., S. O'Brien (2010). "Analysing post-editing performance: correlations with years of translation experience", Proceedings of the 14th Annual Conference of the European Association for Machine Translation, S. Rafael, Francia.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers (2011). "Apertium: a free/open-source platform for rule-based machine translation", Machine Translation 25(2):127-144.

Garcia, I. (2005). "Long term memories: Trados and TM turn 20", The Journal of Specialized Translation 4:18-31.

Guerberof, A. (2009). "Productivity and quality in MT post-editing", Proceedings of the MT Summit XII workshop: beyond translation memories: new tools for translators, Ottawa, Canadá.

Hearne, M., A. Way (2011). "Statistical machine translation: a guide for linguists and translators", Language and Linguistics Compass 5(5):205-226.

Hutchins, W. J., H. L. Somers (1992). "An introduction to machine translation", Academic Press. [Disponible en línea: http://www.hutchinsweb.me.uk/IntroMT-TOC.htm; última visita: 30 de septiembre de 2012]

Koehn, P. (2005). "Europarl: a parallel corpus for statistical machine translation", Proceedings of the Tenth Machine Translation Summit, p. 79-86, Phuket, Tailandia.

Koehn, P. (2010). "Statistical Machine Translation", Cambridge University Press.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst (2007). "Moses: open source toolkit for statistical machine translation", Proceedings of the Annual Meeting of the Association for Computational Linguistics Demo and Poster Sessions, p. 177-180, Praga, República Checa.

Koehn, P., F. J. Och, D. Marcu (2003). "Statistical phrase-based translation", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, p. 48-54, Edmonto, Canadá.

Kranias, L., A. Samiotou (2004). "Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration", Proceedings of the 4th International Conference on Language Resources and Evaluation, p. 331-334, Lisboa, Portugal.

Lagoudaki, E (2008). "The value of machine translation for the professional translator", Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, p. 262-269 Waikiki, EUA.

Och, F. J. (2005). "Statistical machine translation: foundations and recent advances ", Tutorial en Tenth Machine Translation Summit, Phuket, Tailandia.

Penkale, S., R. Haque, S. Dandapat, P. Banerjee, A. K. Srivastava, J. Du, P. Pecina, S. Kumar Naskar, M. L. Forcada, A. Way (2010). "MaTrEx: the DCU MT system for WMT 2010", Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, p.143-148, Uppsala, Suecia.

Phillips, A. B. (2011). "Cunei: open-source machine translation with relevance-based models of each translation instance", Machine Translation 25(2):161-177.

Plitt, M., F. Masselot (2010). "A productivity test of statistical machine translation post-editing in a typical localisation context ", The Prague Bulletin of Mathematical Linguistics 93:7-16.

Rafalovitch, A., R. Dale. (2009). "United Nations General Assembly resolutions: a six-language parallel corpus ", Proceedings of the Twelfth Machine Translation Summit, p. 292-299, Ottawa, Canadá.

Sánchez-Martínez, F., M. L. Forcada (2011). "Free/open-source machine translation: preface", Machine Translation 25(2):83-86.

TAUS (2009). "LSPs in the MT loop: current practices, future requirements", Informe disponible en http://www.translationautomation.com/reports/lsps-in-the-mt-loop-current-practices-future-requirements (última visita: 30 de septiembre de 2012)

Tiedemann, J. (2009). "News from OPUS - a collection of multilingual parallel corpora with tools and interfaces", Recent Advances in Natural Language Processing (vol V), p. 237-248, John Benjamins, Amsterdam/Philadelphia